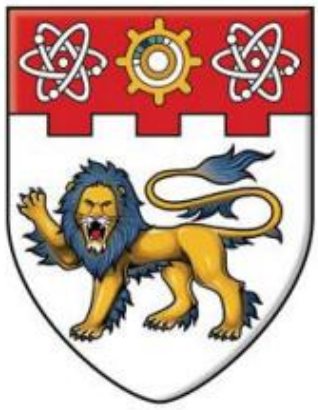

NANYANG TECHNOLOGICAL UNIVERSITY

Building the Foundation Text for Nanyang Technological University – Multilingual Corpus (NTU-MC)

TAN LI LING

SCHOOL OF HUMANITIES AND SOCIAL SCIENCES

2011

Building the Foundation Text for Nanyang Technological University – Multilingual Corpus (NTU-MC)

Liling Tan
Nanyang Technological University
Singapore, Singapore
tanl0087@e.ntu.edu.sg
Matriculation Number: 088355A12

Bachelor Dissertation
Spring, 2011

Project Supervisor: A.Prof Francis Bond

Abstract

The NTU-MC is a multilingual corpus that taps on the availability of multilingual text available in Singapore. The current version of NTU-MC contains a total of ~375,000 words (15,096 sentences) for the NTU-MC in 6 languages (English, Chinese, Japanese, Korean, Indonesian and Vietnamese) from 6 language families (Indo-European, Japonic, Austro-Asiatic, Sino-Tibetan, Austronesian and Korean as a language isolate); all text in English, Chinese, Japanese, Korean and Vietnamese were Part Of Speech (POS) tagged. This project focuses on compiling the foundation text for the NTU-MC and this dissertation describes the motivations, the corpus compilation process and internal and cross-corpora evaluation of the corpus output. The corpus will be made available to the public under the Creative Common – Attribute 3.0 Unported license in Summer 2011.

Assignment Submission Declaration

School of Humanities and Social Sciences

Name:	Tan Li Ling
Matriculation No:	088355A12
Title:	Building the Foundation Text for Nanyang Technological University – Multilingual Corpus (NTU-MC)
Course and Code:	HG499 – Graduation Project
Lecturer/Tutor:	A.Prof Francis Bond
Submission Date:	27 April 2011

Keep a Copy of the Assignment

Please make a copy of your work. If you have submitted your assignment electronically also make a backup copy.

Plagiarism and Collusion

Plagiarism: to use or pass off as one's own, the writings or ideas of another without acknowledging or crediting the source from which the ideas are taken.
Collusion: submitting an assignment, project or report completed by another person and passing it off as one's own (as defined in the *NTU Honour Code*. See www.ntu.edu.sg/sao/home for the University Honour Code and Pledge).

Penalties for Plagiarism and Collusion

The penalties associated with plagiarism exist to reward good academic conduct; those who cheat will be severely punished to reflect the seriousness with which NTU views cheating, and its commitment to academic integrity. Penalties may include: the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment.

Declaration

I declare that this assignment is my own work, unless otherwise referenced, as defined by the NTU policy on plagiarism. I have read the NTU Honour Code and Pledge.
http://www.ntu.edu.sg/home/yclai/

Signed…….. Date ……..

Acknowledgement

This project is made possible by the permission to utilize the data source from *www.yoursingapore.com.sg*. The author of this dissertation thanks *Ms Joan Lee,* New Media Manager of STB, for graciously allowing the usage of the text from the multilingual STB websites. Also, the author's gratitude extends to *Ms Dorothy Cheung*, Public Relations Manager of Sembawang Town Council (SBTC) and *Mr Edrick Chua*, Assistant Director of Corporate Communications from National Environment Agency (NEA) for their permission and aid in provide the data for this project. Though the data from SBTC and NEA is not utilized for the current phase in NTU-MC compilation, it is useful resources for the future extension of the corpus and its toolkits.

Index

1. Corpora and Multilingualism

Corpus linguistics has evolved from the primitive methodology of empirically investigating linguistic hypotheses to a full-blown field in the inter-disciplinary study of computational linguistics. The technique of linguistic data-mining developed into a field of machine-readable corpora building of massive quantity and in multiple languages. Subsequently, this magnitude of corpora building impelled the technical tools required for automatic statistical corpora annotation, queries. Ultimately the availability of these corpora propelled the state of Natural Language Processing (NLP).

As globalization shrinks the world, the appetite for building monolingual English corpora (e.g. Brown corpus by Kucera and Francis, 1967; Australian Corpus of English by Peters, 1987) also advanced to building sizable, automatically annotated multilingual corpora (e.g. Stockholm Multilingual Treebank by Volk and Samuelsson, 2004; European Parliament Proceedings Parallel Corpus by Koehn, 2005).

The linguistic diversity in Singapore is obvious through the visibility and saliency of languages used on public and commercial signs. Such diversity is also reflected in the virtual linguistic landscape, where Singapore based websites are available in multi-languages. These manifestations of societal multilingualism have yet to be tapped on for computerized data collection for linguistic researches. The Nanyang Technological University Multilingual Corpus (NTU-MC) pioneers the initiative to make these multilingual data machine-readable for future NLP tasks. The construction of the NTU-MC is driven by the mantra of empirical and balanced researches where societal linguistic diversity is represented in crystalized form of Languages (i.e. corpora), and using corpora as a baseline to explore linguistic phenomena and to advance NLP techniques.

For this project, the main concern is to build a machine-readable foundation text for the NTU-MC and to encourage corpus linguists to build on the corpus by expanding the size, representation, the layers of annotations and relevant toolkits. The NTU-MC strives to be an ongoing effort to tap on the miasma of ‘exotic’ multilingual texts available in Singapore; ranging from the multilingual sign boards with official languages of Singapore (English, Chinese, Malay, Tamil) to posters, signs and guides targeted towards migrants, expats and tourists (in Indonesian, Japanese, Korean, Vietnamese, Thai, Tagalog, etc.).

1.1 A Brief Introduction to Corpus Linguistics

Corpus linguistics started as a methodology to study linguistic phenomena with the assumption that language is best investigated by analyzing real instances of linguistic production. These samplings (corpora) of real world instances reveals patterns that runs under the radar of linguists relying on their intuition or competence of the language of concern (McEnery and Wilson, 2001). Since the landmark publication of the Kucera and Francis (1967) on 'Computational Analysis of Present-Day American English' based on the Brown corpus, it spawned the trend to build similar corpora and grammatical analysis. For example, the Lancaster-Oslo-Bergen Corpus built by Johansson et al (1986) and the "Corpus-based grammar – the comprehensive Grammar of English" by Quirk et al (1985). Subsequently the race to build machine readable corpora for Natural Language Processing (NLP) tasks commenced as computational linguistics enters the data-driven phase with statistical NLP in the late 1990s (Manning and Schutze, 1999).

The fundamental approach to corpus linguistics is to emulate the real world instances. Empirically, this is achieved by building corpus with authentic (i.e. naturally occurring) data of significant size and coverage of domains that are representative of the language. Nowadays, the original emulative nature of corpora persists but with the additional demand for corpora to be made machine readable for NLP tasks, especially with the increase simplicity of colossal sized corpora (e.g. the Big Web Corpus (BiWeC) is a corpus, currently of 5.5 billion words, with a target size of 20 billion by Pomikalek et al, 2009). Such projects are motivated by the desire to capture linguistics features that occur in rare instances. This is explained by Zipf's (1935) law of word distribution, the probability of rare linguistic feature instances increases proportionally with the size of the corpus, thus it is necessary for corpora to maintain its gianormousity if the purpose of the corpus is to capture the widest range of linguistic features possible in a language (Aston and Burnard, 1998).

The representation aspect of a corpus is crucial in the design of any corpus and central to the purpose of the corpus. To optimally represent the language, there are two sampling techniques that are commonly implemented in corpora building, viz. *proportional sampling* and *stratified sampling* (McEnery and Wilson, 2001). Proportional sampling is the collection of data for the corpus (i.e. text, soundclips, videos, images, etc.) that adheres to the proportion of language use within language community in concern, this implies the statistical measurement of the language production and reception of the population like how

sociological researches sample for human demographics (Biber et al, 1998). Abiding by the proportional sampling philosophy, 90% of the corpus data should be composed of conversation transcripts, since most people spend more time speaking and listening than reading and writing (Sinclair, 2005). Biber et al. (1998) argues that proportional sampling is apt to exemplify a speaker's use of the language in his/her daily usage but it is inadequate to represent language as a whole; they proposed the stratified corpus construction to overcome the vacuous data that may slip the proportional sample. The strata based sampling is likened to slicing the linguistic setting up into domain based strata and samples from each strata were collected to compile a comprehensive representation of the language.

The Nanyang Technological University Multilingual Corpus (NTU-MC) adopts the stratified sampling methodology in data collection and is representative of the multiple languages used in different domains of Singapore language habitus. The current project focuses on the tourism domain of Singapore where multiple foreign languages (English, Chinese, Japanese, Korean, Indonesian and Vietnamese) are used on Singapore Tourism Board's website to entreat the tourism from the countries that speaks the respective languages (refer to Section 2.1 for details on data sources of this project). The project's scope is to build the foundation text for the NTU-MC from the Singapore Tourism Board's webpage *www.yoursingapore.com.* This dissertation describes the motivation to build the corpus (section 1.2), the process of building the corpus (section 2), and evaluation of the corpus compilation (section 3) and concluding remarks on the corpus project (section 4).

1.2 Motivation to build a Multilingual Corpus in Singapore

Singapore multicultural and multilingual society has necessitated the use of parallel text for signboards, public announcements and information dissemination. Tapping on the linguistic diversity in Singapore, NTU Multilingual Corpus aims to compile the parallel multilingual texts that are readily available online and off the streets of Singapore. Our concern is to achieve machine readable text that will be accessible for further Natural Language Processing (NLP) tasks and linguistic researches.

Different from existing multilingual corpora that are constructed from European Languages (e.g. Europarl Corpus built (Koehn, 2005); Balkan SETIMES (Tiedemann, 2009)), NTU-MC represents an array of Asian Languages; providing the linguistic community with a balanced view in corpus linguistics and more diverse data for cross-lingual NLP tasks.

Ideally, the task to achieve the virtualization physical multilingual text should focus on Optical Character Recognition (OCR) technology, the corpus compilation and tagging layers of annotations. To scale the NTU-MC corpus building task to a feasible portion, the project focuses on the tasks of compiling a foundation text and applying scripts and programs to provide basic Part Of Speech (POS) annotation for NTU-MC.

2. Building the Nanyang Technological University Multilingual Corpus (NTU-MC)

The initial proposal was the initiative to build a multilingual corpus with the resources available to the public in Singapore. The NTU-MC took off as an ad-hoc research project from the Linguistic and Multilingual Studies Division of NTU and this project focused on building the foundation text for the corpus within the time frame of a semester.

This section of the dissertation will describe the tasks involved in building the NTU-MC as well as the challenges in the individual task and the corresponding solutions. The description begins with the data source for the NTU-MC and proceeds with the report on the four sub-tasks of the compilation (1) Crawling for data, (2) Cleaning the data, (3) Segmenting the text and (4) POS Tagging the text.

2.1 Multilingual Sources for the corpus

The corpus project was granted the permission to utilize offline texts distributed from the National Environment Agency of Singapore (NEA) and Sembawang Town Council (SBTC). Subsequently, the project was also granted the permission to the websites that are published by Singapore Tourism Board (STB).

Several OCR software (Adobe Acrobat 7, Free-OCR (T-reinhardt.ch, n.d.) and OnlineOCR (2009)) were implemented to convert the resources into machine readable texts but the attempts were unsuccessful. Because of the Latin alphabet bias and the primitive state-of-the-art, OCR systems output English and Malay texts from the scanned documents with an amount of noise; moreover the systems were incapable of processing the Chinese and Tamil texts. Due to the tedious process of computerizing the offline texts manually, the project focused on the electronic texts (the texts from the STB websites) until better OCR technology is available to ease the computerization of offline text.

The corpus project has managed to build the foundation text of the corpus totaling to ~375,000 words[1] (15,096 sentences) in 6 languages (English, Chinese, Japanese, Korean, Indonesian and Vietnamese), from 6 language family trees[2] (Indo-European, Japonic, Austro-Asiatic, Sino-Tibetan, Austronesian and Korean as a language isolate), available on the Singapore Tourism Board's *www.yoursingapore.com* website. All examples used in this dissertation are taken from the NTU-MC.

[1] Excluding punctuations and symbols
[2] Language trees in this dissertation refers to the highest level of language classification from the Ethnologue (Lewis, 2009)

2.2 Constructing the NTU Multilingual Corpus

The NTU Multilingual Corpus is built on a Dell Optiplex 760 Intel server with the Ubuntu 10.10 Lucid Operating System. The corpus is currently hosted under the `IKOMA` server within the Linguistic and Multilingual Studies Division in Nanyang Technological University (NTU) in Singapore; as more clean and annotated text are added to the corpus, plans to host the corpus on open source hosting websites like www.sourceforge.net are in progress. Meanwhile, the NTU-MC is locally accessible on the `IKOMA` server within NTU.

2.2.1 Crawling

Httrack, an open source crawler under the GNU General Public License (Roche, 2007), was used for data-collection in this corpus compilation. The command used in crawling: `httrack http://www,yoursingapore.com -o +*.yoursingapore.com/content/traveller/*/*.html-p1`. This command downloaded the raw HyperText Markup Language (HTML) without the embedded media files (e.g. images, flash files, embedded videos, etc.) from the Uniform Resource Locator (URL) *www.yoursingapore.com/content/traveller*. The other webpages did not have parallel text in languages other than English therefore the project focused compiling on this sub-domain of the website as the foundation text for NTU-MC.

2.2.2 Cleaning

As the markup language used to construct the websites were consistent, a custom-made perl script was created to extract the main body paragraph instead of using the commonly used Condition Random Field (CRF) algorithm (Marek et al., 2007). The HTML files are parsed through the `HTML::TreeBuilder` Perl modules, and only the `<p>…</p>` attributes within the `<div class = paragraph section>…</div>` were extracted. The perl script successfully downloaded the main body text from each webpage and ignored the subtexts that were headers to other pages.

The primary drawback of the extracting the texts bounded by `<p>…</p>` is the inclusion of whitespaces from empty `<p>…</p>` sections. This was solved with secondary whitespace cleaning through `$sed'/^$/d'-i*.txt` bash command. The final problem with whitespaces did not surface until the output files from the alignment tasks were examined; there were non-break spaces (`0xa0`) that were prevalent across the English, Vietnamese and Indonesian data and there were several in the Japanese data too. These `0xa0`

were rampant before the start of the sentence and after the fullstops in the sentence. These `0xa0` were particularly hard to identify because they were invisible to human eyes and were only observable through searching the textfiles for the `0xa0` hexadecimal. A python script was written to search through all extracted text and replace `0xa0` with regular spaces and subsequently all the text that were previously POS-tagged or aligned were re-tagged and re-aligned.

The cleaned data from each language was stored in individual folders, and each text file within the folder contains the extracted main body text from HTML files from *www.yoursingapore.com*. Each line of text in every file encodes the paragraph of one or several sentences. The outputs of the cleaning task are the individual directories (one directory per language) of text files, and each text file contains paragraphs of sentences separated by a newline. All the textfiles were saved in UTF-8 encoding.

2.2.3 Sentence and Word level Segmentation

Different sentential and word level segmentation modules were implemented on the various languages in the corpus. Sentence level segmentation was required for all languages, and word level segmentation was necessary for Japanese, Korean, Chinese and Vietnamese texts.

The nature of word level segmentation tasks in computational linguistics splits sentences up into individual "meaningful units" and these meaningful units are dependent on the philological stance of the individual segmenter and POS tagger programs. The programs chosen to annotate the corpus treats English and Indonesian words as individual tokens separated by whitespaces; and for Japanese, Korean, Vietnamese and Chinese, segments were split based on morphemic or lexical analyzers (not by whitespaces). In this dissertation, the term word and token will be used interchangeably to refer to the individual tokens output by the POS taggers and segmenters.

The English, Korean and Indonesian Texts use the same punctuations and the Natural Language Took Kit (NTLK) `sent_tokenize` module (Bird et al, 2009) was sufficient to segment the English, Korean and Indonesian text. The `sent_tokenize` program uses stop punctuations (i.e. ! ? .) to identify the end of the sentence and also it has an en suite function to differentiate between the fullstop used at tend of a sentence and within a website stated in a

sentence. For example, the `sent_tokenize` correctly segments a sentence with website, which contains multiple fullstops, as one sentence:

'for their reviews and ratings, please purchase the latest makansutra singapore 2011 edition or visit www.makansutra.com. '

The multi-byte Chinese and Japanese sentences were separated by the same sets of ！ ？ 。 punctuations. Thus the `nltk.RegexpTokenizer(u'[^ ！ ？ 。 ]*[！ ？ 。]')` was used to segment the Chinese and Japanese sentences. The Japanese regex has a minor tweak from the common `nltk.RegexpTokenizer(u'[^「」！ ？ 。]*[！ ？ 。]')` , as recommended by the Hagiwara's (2010) Japanese chapter of the Japanese 'Natural Language Processing with python「入門 自然言語処理」' (Bird et al, 2010). The tweak was necessary to include non-sentence phrases bounded by「…」brackets. Normally the Japanese「」brackets would have an individual sentence within the bracket, the text from *www.yoursingapore.com* used the 「」differently by embedding not only sentence but also proper names (e.g. 「マリーナ貯水池」*marina chosuichi* "Marina Reservoir"; 「スターバックス」*sutabakkusu* "Starbucks") or loan phrases(e.g.「三歩一拝」*san ho ichi hai* "three step a bow"- a Chinese Buddhism term; 「ハラール」 *harau* "halal"; 「カルーセル」*karuseru* "carousal") .

For the Japanese and Korean word level segmentation, the segmenter is incorporated into the POS-taggers that this corpus project is using. Similar to Japanese and Korean, the Chinese characters do not require whitespace as a word delimiter. Thus it is necessary to split the characters up into sensible words before POS-tagging the words. The Stanford Chinese word segmenter was used to segment the Chinese sentences in this corpus. The Stanford segmenter relies on linear-chain conditional random filed (CRF) model to segment the words, it identifies character N-grams that matches the ones in the Chinese lexicon (i.e. dictionary word) and treat the initial character of the match as a binary 1 and the subsequent characters in the N-gram as 0. Grouping each 1*[^1] binary regex will result in the individual word segment (refer to example below). Although Stanford segmenter also uses morphological and character features to segment, it does mis-segments unknown words that the program has not encountered before. The mis-segments were mostly local street names that were transliterated from English to Chinese. The reported accuracy of the Stanford Chinese segmenter is a F-score of 94.7% on the Academia Sinica corpus (Tseng et al, 2005). Below is an example of

how the Stanford Chinese word segmenter segment the corpus data, the correct segment for should be 乌节路 *wujie lu* "Orchard road" instead of 乌 节路 *wu jielu* "black joint-road". (These topological mis-segments are left uncorrected for the current version of NTU-MC because a machine readable lexicon of Singapore road names in the languages available in NTU-MC is unavailable. Post-annotation cleanup will be done before the public release of NTU-MC):

(1)	*从*	*乌*	*节路*	*的*	*东陵*	*购物*	*中心*	[Chinese segmented text]
	1	1	10	1	10	10	10	[CRF binaries]
	from	black	joint-road	of	Tanglin	shopping	centre	

'From the Tanglin Shopping Centre of Orchard Road'

For the Vietnamese sentence and word segmentation task, JVnTextPro (Nguyen and Phan, 2007) was used in processing the Vietnamese data in NTU-MC. It is an open source Java based tool, which is a one-stop program to segment a sentence, and segment the words and finally POS tag the data. Using simple regex and Vietnamese sentence structure rules, the JVnTextPro is able to segment the paragraphs from the cleaned data into sentences. Then the JVnSegmenter based on Conditional CRF (FlexCRFs) algorithm (Nguyen et al, 2006) was used to process the word-level segmentation. The JVnSegmenter reported an accuracy of 94.05% F-score based on segmentation of newspaper articles. Although Vietnamese words are separated by whitespaces in the orthography, sometimes two "words" separated by whitespace are supposed to mean a single thing. For example, the Vietnamese word 'quốc tế' mean international but the individual "word" separated by the space does have its meaning ('quốc' means country and 'tế' means to run). Without the correct word boundary the word quốc tế would have been tagged quốc, a noun and tế, a verb. The correctly segmented words boundaries were marked by whitespaces and words like quốc tế were identified with an underscore between (i.e. quốc_tế). A sample of the word level segmented Vietnamese data is presented below; there are some mis-segment like hàng đầu 'leading' were segmented as hàng 'found' and đầu 'head'. The other problem is the disparity between the English version and the Vietnamese version of the webpage, there are much more details in the English webpage and also more updated where the Vietnamese site stated that 'the shops will be opened soon' and the English site infers that the shops are already opened. This will invariably cause problems in the alignment task.

Word level segmented Vietnamese text from NTU-MC (`marina-bay-sands.txt`):

(2)	*Một*	*loạt*	*các*	*cửa_hàng*	*cao_cấp*	*và*	[Vietnamese]
	one	series	all	shops	advance[3]	and	
	tinh_tế	*sẽ*	*sớm*	*được*	*khai_trương*	*với*	
	exquisite	will	soon	be	opening	with	
	những	*thương_hiệu*	*quốc_tế*	*hang*	*đầu*	.	
	these	brand	inter-national	found	head	.	

'A variety of new and exquisite shops will soon be launched with the leading international brand.'

The English version of `marina-bay-sands.txt` in the NTU-MC:

'With a wide array of high-end boutiques alongside niche designer labels at the shoppes at Marina Bay Sands , featuring top international brands such as Louis Vuitton located in a "floating" crystal pavilion , shoppers will surely be spoilt for choice.'

2.2.4 Part of Speech (POS) Tagging

For POS-tagging tasks, different programs were implemented for the individual languages and the different languages were tagged with a set of different POS tag depending on the program. All data except the Indonesian texts were POS-tagged, this is due to the lack of an open source POS-tagger for Bahasa Indonesian. All the tagged output was formatted into the Corpus Work Bench (CWB) verticalized text format with eXtensible Markup Language (XML) tags to encode the start and end of a sentence (i.e. `<s>…</s>`), and each word/token separated by a newline and each line contains a word and its POS tag separated by a tab. Table 1 presents a brief summary of the sentence segmentation and POS-tagging task for the corpus compilation.

The HunPos tagger (Halacsy et al, 2007) was implemented to apply the POS annotation to the English texts. HunPos is an open source Hidden Markov Model (HMM) based trigram tagger, emulating the free but not open TnT tagger (Brants, 2000). The pre-trained Wall Street Journal English (`en_wsj.model`) model was used with the HunPos tagger to tag the English data in NTU-MC. The HunPos tagger reported an accuracy of 96.88% seen words (i.e. words that the program's model has seen when the program was trained) and 86.13% for unseen words. The greatest advantage against other English taggers is its speed; it

is "of order of magnitude faster than the current generation of SVM and MaxEnt" taggers (Halacsy et al, 2007). Although the current NTU-MC is small in size, speed is always an advantage to any computing task. The other incentive is its ease of usage; the HunPos, originally written in OCaml, was recently incorporated into the python system, this eludes the task to learn another computing language to implement an annotation to a corpus. The Penn Treebank II tags (Mitchell et al, 1993) were used by the HunPos tagger (refer to appendix for the complete list of tagset).

Language	Sentence Segmenter	Word Segmenter	POS-tagger	POS Accuracy	Encoding	Tagset
English	NLTK sent_tokenize	whitespaces	HunPos	96.58%	ISO-8859-1	PennTreeBank Tagset
Japanese	NLTK RegexpTokenizer	MeCab	MeCab	96.75 – 97.66%	UTF-8	IPAdic Tagset
Korean	NLTK RegexpTokenizer	POSTAG/ Sejong	POSTAG/ Sejong	90.7%	EUC-KR	Sejong Tagset
Vietnamese	JVnTextPro	JVnSegmenter	JVnTagger	93.32%	UTF-8	VSLP Tagset
Chinese	NLTK sent_tokenize	Stanford Segmenter	Stanford POS-tagger	93.65%	UTF-8	Penn Chinese TreeBank Tagset
Indonesian	NLTK sent_tokenize	whitespaces	-	-	-	-

Table 1: Summary of Segmentation and POS tagger used in NTU-MC

The word segmented Japanese data were tagged by the MeCab tagger (Kudo et al, 2004). MeCab is an open source morphological analyser developed using the Conditional Random Fields (CRF) bi-gram model. It has an improved performance, in both speed and accuracy, than its predecessor ChaSen (HMM model based) tagger (Matsumoto et al, 1999). The MeCab tagger was used with the `-Ochasen` model, which was trained by the ChaSen tagger. Similar to HunPos, the MeCab tagger has an Application Programming Interface (API) in the python language and is freely available through the Ubuntu NLP repository. The MeCab morpheme analyser reported an overall F-score (harmonic precision/recall score) of 96.75% on the Kyoto University Corpus ver. 2.0 and 97.66% on the Real World Computing Partnership Text Corpus (Kudo et al, 2004). Different from the other POS-tagger used in this project, the MeCab morphological analyser provided more than a layer of POS annotations; MeCab output adheres to the IPADIC 2.7.0 standards (Asahara and Matsumoto, 2003) which are compatible with the ChaSen tagger.

The POSTech TAGger –Korean (POSTAG/Sejong) was used to tag the Korean text in NTU-MC. As an agglutinative language, POSTAG/Sejong tagged the tokens at a morpheme

level rather than a word level. POSTAG/Sejong used a hybrid tagging model that first made used of Korean morpheme pattern dictionary to segment the sentences into morpheme and then it lumped the related morphemes onto the root morpheme to form a token. After that, each token is tagged based on the statistical Viterbi algorithm (Forney, 1973), and finally a posteriori error-correction was ran through a Brill transformation remapping (Brill, 1995) to correct the mis-tagged morpheme from the Viterbi algorithm. The POSTAG/Sejong was reported to achieve an overall 97% tagging accuracy with data containing 10% unseen words (Lee et al, 2002). The POSTAG/Sejong was a closed (including the model used by POSTAG/Sejong) but it is free to use, hence the drawback of the POSTAG/Sejong is the need to pipe the EUC-KR output from POSTAG/Sejong into python to process and format the output into UTF-8 CWB format. The other challenge was the availability of POSTAG/Sejong, it was only available on Microsoft Windows OS but the server was running on a Unix based Ubuntu box. Short of processing the data, with another computer, a shell script was made to implement the POSTAG/Sejong on Wine (Windows emulator). The custom tagset with 41 tags was used by POSTAG/Sejong to suit the Korean morpheme (refer to appendix for POSTAG/Sejong tagset).

The Stanford Chinese POS tagger (Tseng et al, 2005) was use to tag the Chinese data. The Stanford Chinese POS tagger was built on the MaxEnt Markov Model (MEMM) by calculating the probabilities of state transition (Low et al, 2005). The model used for Stanford Chinese tagger (`chinese.tagger`) model was applied to the Stanford MEMM tagger to tag the Chinese text. The Stanford Chinese POS tagger reports an overall accuracy of 93.65% and 84.84 for unseen words. The choice of POS tagger was based on its state-of-art accuracy. Although the setback is the disability to use the Java-based Stanford tagger concurrently with the python scripts used for the other POS tagging task in this project, a shell script was written to pipe (i.e. technical term for outputting data from Unix terminal) the output of the tagger to format into verticalized text format. The Chinese Penn Treebank tagset (Xia et al, 2000) were used by the Stanford tagger.

For the Vietnamese data, the JVnTagger (part of the JVnTextPro tool) was implemented. The JVnTagger is based on the both CRF and MaxEnt models. Both models achieved an approximate 93% (F-score), but the CRF model yielded 93.45% on the 10,000 words Vietnaemse TreeBank (VTB) while the MaxEnt model scored 93.32% on the 20,000 words VTB (Nguyen et al, 2010). For this project, the MaxEnt (`maxent`) model was chosen

purely based on the fact that it has encountered more words in the training process. Though open sourced, the tagger was coded in Java, so the output was pipped into python to format the text into verticalized text format. The tagset used in by JVnTextpro sets the standards as they pioneered the VLSP project (2006-2010) to "building basic resources and tools for Vietnamese language and speech processing", a five year long project from 2006 – 2010.

The primary issues with multilingual corpus POS annotation is the difference in encoding of the sources and the encoding that the POS tagger accepts as input and produce as output. To standardize the encoding for NTU-MC, the widely accepted UTF-8 is the designated encoding for NTU-MC. There were no issues from the sources, as all the textfiles were saved as UTF-8 format after the cleaning process. However when the UTF-8 encoding is fed into the English (HunPos) and the Korean (POSTAG/SEJONG) tagger, the encoding needs to be changed to the respective encoding that the tagger accepts. Such encoding problems are fairly common for programmers dealing with European (ISO-8859-1) and Chinese characters (UTF-8), thus there are resolution device within the python programming languages.

For the English data, the python decode function was used during the POS-tagging to decode UTF-8 and encode to ISO-8859-1, and the reverse encoding were used to save the HunPos output into UTF-8. However the less common encoding issues surface when using the POSTAG/SEJONG program for the Korean data. The POSTAG/SEJONG only tags data in EUC-KR encoding, most of the characters were successfully encoded and decoded during the POS-tagging process but there was no direct way of mapping the UTF-8 to the EUC-KR encoding for certain characters. The – , é and © Unicode characters were replaced with - , e and (C) during the POS-tagging task for the Korean texts.

Table 2 presents a sample output of the POS-tagged data of the extracted and cleaned text originally from *http://www.yoursingapore.com/content/traveller/en/browse/see-and-do/nightlife/dance-clubs/zouk.html*. This is the first sentence from the main text in the webpage in five languages. The sample includes the individual words/tokens and their respective POS. The other non-POS annotations that were in the Japanese data were excluded in the table for comparison with the other languages.

English		Japanese		Korean		Vietnamese		Chinese	
Token	**POS**	**Token**	**POS**	**Token**	**POS**	**Token**	**POS**	**Token**	**POS**
<s>		<s>		<s>		<s>		<s>	
If	IN	シンガポール	名詞-固有名詞-地域-国	싱가포르	NNP	Nếu	C	如果	CS
you	PRP	で	助詞-格助詞-一般	에서	JKB	bạn	N	您	PN
only	RB	一つ	名詞-一般	클럽	NNP	chỉ	R	在	P
have	VBP	の	助詞-連体化	한	NNP	có	V	新加坡	NR
time	NN	クラブ	名詞-一般	군데	NNB	thời_gian	N	只	AD
for	IN	に	助詞-格助詞-一般	밖에	JX	ghé	V	能	VV
one	CD	しか	助詞-係助詞	가	VV	thăm	V	前往	VV
club	NN	行く	動詞-自立	ㄹ	ETM	một	M	一	CD
in	IN	時間	名詞-副詞可能	시간	NNG	câu_lạc_bộ	N	间	M
Singapore	NN	が	助詞-格助詞-一般	이	JKS	ở	E	俱乐部	NN
,	,	なかっ	形容詞-自立	없	VA	Singapore	Np	,	PU
then	RB	た	助動詞	다면	EC	,	,	祖卡	NN
it	PRP	と	助詞-格助詞-引用	,	SP	hãy	R	酒吧	NN
simply	RB	し	動詞-自立	Zouk	SL	đến	V	必然	AD
has	VBZ	たら	助動詞	를	JKO	Zouk	Np	是	VC
to	TO	、	記号-読点	선택	NNG	.	.	您	PN
be	VB	間違い	名詞-ナイ形容詞語幹	하	XSV	</s>		的	DEG
zouk.	JJ	なく	助動詞	시	EP			不二	JJ
</s>		、	記号-読点	어요	EF			选择	NN
		この	連体詞	.	SF			。	PU
		ズーク	名詞-一般	</s>				</s>	
		に	助詞-格助詞-一般						
		行く	動詞-自立						
		べき	助動詞						
		です	助動詞						
		。	記号-句点						
		</s>							

Table 2: Sample POS-tagged Outputs of the NTU-MC

3. Evaluating the NTU Multilingual Corpus

As of the current state of NTU-MC, the foundation text from *www.yoursingapore.com* were extracted, cleaned and POS tagged. The corpus will be made available to the public possibly in summer 2011. This section will describe (1) the availability of the corpus, (2) the output from this corpus project, (3) an evaluation of the corpus internally and (4) with comparable corpora.

3.1 Availability of the Corpus

For a corpus to be a valuable resource, it must be both useful and accessible (Ishida et al. 2006). The current NTU-MC is available under copyrighted under the Creative Common (CC) Attribution 3.0 Unported License. Users of the corpus are able to share (i.e. copy, distribute and transmit) and remix (i.e. to adapt) the corpus under the condition of attributing the work to the NTU-MC project (a summary of the license is attached in the Appendix). The owners of the source data (Singapore Tourism Board, Sembawang Town Council, National Environmental Agency) were notified of the license and had agreed upon allowing the use of the data for this project.

3.2 Corpus Output

The NTU-MC project compiled a foundation text of ~375,000 words[3] (15,096 sentences) for the NTU-MC in 6 languages from 6 language family trees, available on the Singapore Tourism Board's *www.yoursingapore.com* website. The breakdown of the sentences is as followed (the language code corresponds to language code used by the website; the no. of tokens excludes punctuations and symbols):

Language	Language code	No. of Tokens	No. of Sentences	POS tagged	Language Family Trees
English	en	76,339	3,255	✓	Indo-European
Japanese	ja	72,797	2,648	✓	Japonic
Korean	ko	67,341	2,407	✓	Language Isolate
Vietnamese	vi	56,535	2,236	✓	Austro-Asiatic
Chinese	zh	52,047	2,365	✓	Sino-Tibetan
Indonesian	id	50,315	2,185	✗	Austronesian
Total		375,374 words	15,096 sentences		6 Language Family Trees

Table 3: A Summary of the Corpus Output

[3] The total number of tokens including punctuations and symbols is 415,222

Table 4 proposes a coarse-grained POS tagset that can ease comparison of POS across languages. Though appreciative of the unique nature of each language, a standardized cross-lingual POS tagset would be advantageous for traditional contrastive linguistics as well as cross-lingual computational tasks (the specific count for each individual POS tags defined by the taggers is available in the respective tagsets presented appendix).

	English	**Japanese**	**Korean**	**Vietnamese**	**Chinese**
Affixes	-	接頭詞	XPN,XS,XSA,XSB,XSN,XSV	-	-
Adjectival	JJ,JJR,JJS	形容詞	VA	A	VA
Adnominal	-	連体詞	ETM,MM	-	-
Adverbials	RB,RBR,RBS	副詞	MAG,MAJ	R	AD
Conjunctions	CC	接続詞	EC	C	CC,CD
Foreign/Loan	FW	名詞-固有名詞-人名-一般	SH,SL	B	FW
Interjections	UH	感動詞,フィラー,その他-間投	IC	I	IJ,PP
Nominal	NN,NNS,NNP,NNPS,POS,PDT,DT	名詞	ETN,NNB,NNG,NNP	A,N,Ny,Y,Nc,Np,Nu,Nb	JJ,NN,NR,NT,DT,DEG
Numeral	CD,LS	名詞-数	NR,SN	M	CD,OD
Particles	RP	助詞	EF,EP,JC,JKB,JKV,JKC,JKG,JKO,JKQ,JKS,JX	-	MSP,SP
Preposition /Locational	IN,TO	-	-	E	LC
Pronominal	PRP,PRP$	名詞-代名	NP	P	PN
Punctuations	SYM	記号	SE,SF,SO,SP,SS,SW	Mrk	PU
Question	WP,WP$,WRB	-	-	-	-
Verbal	VB,VBD,VBG,VBN,VBP,VBZ	動詞,助動詞	VCN,VCP,VV,VX,XR	V	AS,VC,VE,VV,DER,DEV
Others (Language specific POS)	-	-	-	X	SB,BA,DEC,ETC,LB,ON

Table 4: Coarse-grained POS Grouping

Other than the corpus text output, several scripts were created in the process of the corpus compilation. These were simple HTML extraction, whitespace cleaning and string manipulation tools, scripted in perl, python or shell. However these scripts were specific to the texts in this corpus; they are applicable to other corpora compilation projects but modification needs to be made. The most valuable script contribution from the NTU-MC is the shell script that implements the POSTAG/Sejong tagger on Unix machines. The free but not open Sejong tagger was only dependent on a Korean Microsoft Windows OS platform. The Sejong-Shell script allows the POSTAG/Sejong to be implemented in Unix based OS through the WINdows Emulator (the short shell script is attached in the appendix).

3.3 Evaluating the POS annotation

The `fish-head-curry.txt` from the NTU-MC was selected at random[4] for human annotators to verify the POS-taggers' accuracy. For each language, either a native speaker of the language (self-reported by the human annotator) or a second language learner of the language was assigned to tag the selected files. It took on average 1.5 hours for verify the tag in the one text.

All annotators were briefed on the tagsets used by the respective taggers before the annotation. The initial evaluation approach was for all annotators to tag the text from scratch; however the lack of in-depth knowledge about the tagsets deters the human annotators to use sophisticated tags. For example, the positive copula (VCP) from POSTAG/Sejong tagger, the human annotator tagged it as a simple verb (VV). Thus the methodology was modified and it required the human annotator to check the tagged data and replace the tag when a tagged token was deemed mis-tagged. The annotators were also instructed to tag the tokens with an asterisk (*) if the tokens were mis-segmented. In such a case, accuracy of the human annotation might be primed by what the POS tagger had tagged but it does reflect the human threshold for accepting a machine tagged text. Therefore the human verifications were not treated as the "gold standard" but an inter-annotation agreement (IAA) score was derived from the annotators' identification of the mis-segmented and mis-tagged tokens. The number of tokens used in the calculations excludes punctuations and both number of mis-segments and mis-tagged contributes to the disagreement score.

[4] The first filename from `$ls |sort -R|tail -$N` Unix terminal command

For the Japanese POS evaluation, there was no available native or second language learners who are familiar with the ipadic POS. Thus a different POS tagger, ChaSen morphemic analyzer, was implemented to calculate IAA. Technically, MeCab is a more advance version of ChaSen but the IAA between the two POS tagger does provide an error bar to the reported MeCab accuracy. Both programs uses the ipadic POS, but the noticeable difference is that ChaSen is more conservation with the POS tag unknown words; ChaSen applied the 未知語 *michigo* "unknown word" tag to tokens that the program had not seen before (e.g. フィッシュヘッドカレー *fisshuheddokare* "fish head curry" or イカンメラ *ikanmera* "ikan merah") but MeCab forces the closest fit POS to the unknown tokens. There were 12 instances of these MeCab's forced tagging cases within the fish-head-curry text, but they were all tagged correctly as nouns. Thus the IAA calculation between ChaSen and MeCab does not include these 12 instances as mis-tagged, but the other differences in POS tags were counted as MeCab's mistag.

Language	No. of tokens	No. of sentences	No. of mis-segments	No. of mis-tagged	IAA	Reported POS tagger accuracy
English	235	7	-	18	**92.23%**	**96.58%** (Halacsy et al, 2009)
*Japanese	293	14	3	8	**96.25%**	**96.75-97.66 %** (Kudo et al, 2004)
Korean	374	14	44	27	**81.02%**	**90.7%** (Lee et al, 2002)
Vietnamese	225	7	14	10	**89.33%**	**93.32%** (Nguyen et al, 2010)
Chinese	249	9	19	16	**85.94%**	**93.65%** (Tseng et al, 2005)

Table 5: Summary of POS Annotation Evaluation

Table 5 above presented a summary of the POS-tag evaluations. The IAA score calculation was adopted from the IAA standards for the semantic annotation in the CLEF corpus (Roberts et al, 2008). The IAA is measured as such:

non-matches = no. of mis-segment + no. of mis-tagged

matches = no. of tokens – non-matches

IAA = matches / (matches + non-matches) * 100%

The IAA reported in table 5 serves as a gauge, an error bar, of the reported accuracy reported by the individual taggers (e.g. the HunPos tagger reflects an overall 95.50% (92.23/96.58 *100%) accuracy when tagging the NTU-MC).

3.4 Comparable Corpora

In terms of relevance of the current foundation text in NTU-MC, it is similar to the written section of the International Corpora of English – Singapore (ICE-SIN); i.e. ~400,000 words out of the full ICE-SIN of 1 million words (Nihilani, 1992; Ooi, 1997). The ICE-SIN was built upon written text from printed materials (e.g. academic writing, news reports, editorials, etc.) and non-printed materials (e.g. correspondences, student essays, exam scripts, etc.). However, the ICE-SIN lacks written resources from the web of Singaporean internet domain because ICE-SIN was compiled in the 1990s, before the explosion of internet materials. Technically, the English portion of the foundation text built in this project could be imported into ICE-SIN since the texts are applicable for both corpora. Moving beyond English, the current NTU-MC represents a stratum of the written multilingual texts in Singapore, specifically the domain of Singapore websites targeted at attracting tourist.

As compared to other multilingual corpora, the present NTU-MC is lacking in size; NTU-MC has an average ~66,000 words/tokens per language. With the current state-of-art NLP techniques, a web crawled corpora should be reaching at least the size of ukWaC (~2 billion words; Ferraresi et al, 2009) or the Wiki Corpora and Web Corpora from the Corpus Factory project (~0.2 – 149 million words (varies for different language data); Kilgarriff et al, 2010). However, the project has met the purpose of building the foundation text for the corpus and the aim of the NTU-MC is to move beyond the machine-readable text and tap on existing multilingual text from the local linguistic landscape. And the unique linguistic landscape of Singapore provided NTU-MC with data from 6 different language families; such diversity in multilingual data opens up new possibilities for NLP tasks that rely on cross-lingual training (e.g. machine translation, cross-lingual word sense disambiguation, etc.).

As much as the output of the data (i.e. corpus size, corpus annotations, corpus usage, etc) is important, the philosophical motive of corpus building is also crucial in prospective academic researches. The OPUS corpora had pioneered the movement to tap on free and open multilingual data from the communities of open source enthusiasts (Tiedemann, 2009). The Europarl Corpus (part of the OPUS project) ingeniously crawled the readily available multilingual data from the translated European Union parliament documents (Koehn, 2005). The NTU-MC aims to draw on the local linguistic diversity and virtualize physical multilingual data that is readily available.

Table 6 summarized the cross-corpora comparison made in this section of the dissertation in terms of size, representation, motives, number of languages and availability of different annotations. The statistics reported in Table 5 are based on the cited papers.

	NTU-MC	**ICE-SIN**	**ukWaC**	**Wiki/Web corpora**	**Europarl (part of OPUS)**	**OPUS**
Size	~ 375,000	~ 1 million (Ooi, 1997)	~ 2 billion (Ferraresi et al, 2009)	~0.2 – 149 million (Kilgarriff et al, 2010)	~ 30 million (Koehn, 2005)	Ever-growing (Tiedemann, 2009)
Represent-ation	Singapore Multilingual Linguistic Landscape	Singapore English Usage	British English from online data within .uk domains	Multilingual data available from seed selected web queries and Wikipedia	Official proceedings from European Parliament	Open source data available online
Motive	Virtualize readily available multilingual resources into machine readable text	Achieve comparable world-wide English corpora	To introduce the Web as Corpus Initiative to the NLP community	Standardize a method to build massive corpora in different languages	Extract translated European Parliament proceedings for Machine Translation training	Use readily available open source data online to build parallel corpora
Availability	CC-Attribute 3.0 Unported	Free with ICE-SIN license	Free	Build it Yourself (BIY) from Wikipedia dumps	Open source	Open source
No. of Language Families[5]	6	1	1	5	2	Ever-growing
No. of Languages	6	1	1	8	11	Ever-growing
POS tag	✓	✗	✓	✗	✓	Depends on sub-corpora
Align-ment	✗	✗	✗	✗	✓	Depends on sub-corpora
Syntactic Parse	✗	✗	✗	✗	✗	Depends on sub-corpora

Table 6: Cross-corpora comparison

[5] Language trees in this dissertation refers to the highest level of language classification from the Ethnologue (Lewis, 2009)

4. Tasks in Progress

The NTU-MC is an ongoing effort to add content, layers of annotation and usability as it continues to make multilingual resources machine readable for NLP tasks. Currently, the project is progressing with the corpus alignment task before the corpus will be made publicly available for the NLP community. The other task in progress is to allow user-friendly queries on NTU-MC through the Corpus Query Processor web (CQPweb) interface.

The current NTU-MC contains parallel texts that are not ready for the alignment task because of the difference in the number of sentences across the textfiles in the various language of NTU-MC. The project remains undecided on whether to align the text manually or to implement sentence level alignment algorithm such as HunAlign (Varga et al, 2005) or Gale-Church algorithm (Gale and Church, 1993). For word level alignment, this project will utilize the mGiza toolkit, a multi-threaded alignment software based on its predecessor - the Giza toolkit (Al-Onaizan et al., 1999; Och, 2000), to map the sentence pairs from a source language to a target language.

The CQPweb is based on the IMS Corpus WorkBench (CWB) system (Evert, 2008). It provides an end-user interface that can allows researchers to query the NTU-MC for corpus linguistics researches on particular language phenomena. At present, the CQPweb interface is installed and the data in the current NTU-MC are also mounted onto the CQPweb platform (hosted on `IKOMA` server). The queries can be made to investigate concordances, collocations and word frequencies. The CQPweb is presently under closed-testing and is limited to local access within NTU; public access to NTU-MC's CQPweb will be available concurrently with the public release of NTU-MC. Though the current amount of text in the NTU-MC is limited, the interface and the corpus can still be used as a teaching material for corpus linguistics.

Figure 1: NTU-MC CQPweb Mainpage

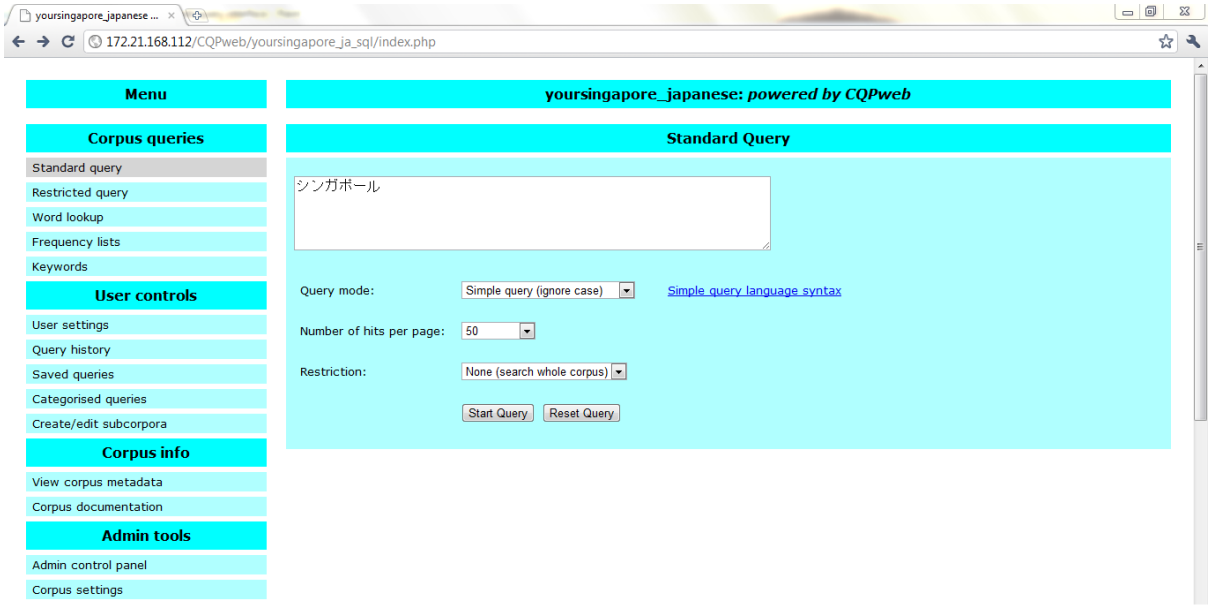

Figure 2: NTU-MC CQPweb Query

Figure 3: NTU-MC CQPweb Sample Query Result

5. Prospective Work on the NTU-MC

The immediate add-on to the project would be to expand the corpus size and representation. As compared to other major corpus (e.g. German Reference Corpus (Kupietz et al, 2010); British National Corpus, 2007), this project is rather small in size but it achieved the purpose of tapping societal linguistic diversity for NLP data. This project urges future research to continue to draw precious NLP data through readily available yet untapped multilingual data; to make multilingual texts within the linguistic landscapes machine readable and to tap on unexploited parallel texts online for corpus compilation.

The other track to future research will be to exploit the corpus for NLP researches once it has matured in size. As opposed to other multilingual corpora that often presents data in languages of the European language families, the concoction of the various languages from an array of language families represented in NTU-MC can provide an interesting view as to how "exotic" cross-lingual (e.g. Vietnamese – Indonesia language pairs) training can advance the current state-of-art NLP tasks such as machine translation and cross-lingual word sense disambiguation. The potential of this 'exotic' concoction in NTU-MC can be accentuated by adding another layer of semantic annotation. A sense tagging task can be implemented manually or automatically (Navigli, 2006), the respective WordNets can be used for the languages[6]. The interest to build the Asian WordNet (Robkop et al, 2010) compliments the future works on sense annotating the NTU-MC and impeccably advancing cross-lingual NLP tasks.

[6] English WordNet; FrameNet (Miller et al, 1993; Fillmore and Baker, 2010), Japanese WordNet (Bond et al, 2009), Korean WordNet (Chagnaa et al, 2007), Viet WordNet (Ho and Nguyen, n.d.), Chinese WordNet (Huang et al, 2010) Indonesian WordNet (Putra et al, 2008)

6. Conclusion

This project has achieved the aim to compile the first ~375,000 tokens (15,096 sentences in 6 languages from 6 language family trees) as the foundation text of the NTU Multilingual Corpus. The NTU-MC will be useful for natural language research because of its diverse multilingual nature. The ultimate goal of the NTU-MC is to tap on available multilingual resources in reality and the virtual world and make them machine readable for NLP tasks.

By bridging the gap between what is machine-readable and what is available, it will inevitably better the representation of the corpus. By progressively extending the NTU-MC with multiple layers of annotation, it will certainly expand the scope of the usage and become a better corpus for general or computational linguistics resources. By building corpora of more diverse cross-lingual nature, it pushes the state-of-the-art NLP techniques through more robust cross-lingual training (Matsumoto et al., 1993).

References

Yaser Al-Onaizan, David Purdy, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Noah A. Smith, Franz Josef Och, and David Yarowsky. 1999. Statistical machine translation. In *Final report, Center for Language and Speech Processing.* John Hopkins University.

浅原正幸 (Masayuki Asahara) and 松本裕治 (Yuji Matsumoto). 2003. *Ipadic version 2.7.0 ユーザーズマニュアル (User's Manual).* Nara Institute of Science and Technology 奈良先端科学技術大学院大学.

Guy Aston and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Steven Bird, Ewan Klein, Edward Loper. 萩原正人 (Masato Hagiwara), 中山敬広 (Takahiro Nakayama) and 水野貴明(Takaaki Mizuno) (translation). 2010. *入門 自然言語処理 (Natural Language Processing with Python)*. O'Reilly, Japan.

Steven Bird, Klein Ewan, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009, Singapore*.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000), Seattle, WA.*

Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics, 21,* pp.543-565.

Altangerel Chagnaa, Choe Ho-Seop, Cheol-Young Ock and Hwa-Mook Yoon. 2007. On the Evaluation of Korean WordNet. In *Václav Matousek and Pavel Mautner (Eds.): TSD 2007. LNCS (LNAI),* vol. 4629, pp.123-130. Springer, Heidelberg.

Constitution of the Republic of Singapore. pt XIII, art. 153A. Official languages and national language. *Retrieved on 13 April 2011 from http://statutes.agc.gov.sg/non_version/cgi-bin/cgi_getdata.pl?actno=1999-REVED-CONST&doctitle=CONSTITUTION%20OF%20THE%20REPUBLIC%20OF%20SINGAPORE%0a&date=latest&method=part&sl=1&segid=931158661-003585#931158661-003601*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning*, vol. 20, pp.273-297.

Yariv Ephraim and Neri Merhav. 2002. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48 (6): pp.1518–1569.

Stefan Evert. 2008. Inside the IMS Corpus Workbench – Presentation at the IULA, Universitat Pompeu Fabra, Barcelona, Spain. *Retrieved on 7 Jan 2011 from http://cwb.sourceforge.net/doc_links.php*

Adriano Ferraresi, Marco Baroni, Silvia Bernardini, Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3): pp.209-226

Charles J. Fillmore and Collin Baker. 2010. A Frame Approach to Semantic Analysis. In *The Oxford Handbook of Linguistic Analysis, Bernd Heine and Heiko Narrog.*, pp.313-339. Oxford: Oxford University Press.

George D. Forney, Jr. 1973. "The Viterbi algorithm". In *Proceedings of the IEEE*, 61 (3): pp.268–278.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19: pp.75–102.

Gary Geunbae Lee, Jeongwon Cha, Jong-Hyeok Lee. 2002. Syllable pattern-based unknown morpheme segmentation and estimation for hybrid part-of-speech tagging of Korean. *Computational Linguistics*, 28(1). pp 53-70.

Silviu Guiasu and Abe Shenitzer. 1985. The principle of maximum entropy. *The Mathematical Intelligencer*, 7(1), pp.42-48.

Péter Halácsy, András Kornai, Csaba Oravecz. 2007. HunPos - an open source trigram tagger In *Proceedings of the 45th Annual Meeting of the Association for Computational*

Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic, pp.209--212.

Andrew Hardie. forthcoming. "CQPweb - combining power, flexibility and usability in a corpus analysis tool". *Retrieved on 7 Jan 2011 from http://cqpweb.lancs.ac.uk/*

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2008. Chinese Wordnet: Design, Implementation, and Application of an Infrastructure for Cross-lingual Knowledge Processing. In *Proceedings of the 9th Chinese Lexical Semantics Workshop.*

Ngoc-Duc Ho and Thi-Thao Nguyen. n.d. Towards Building a WordNet for Vietnamese. *Retreived from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.68.2916*

Toru Ishida. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp.96-100, keynote.

Stig Johansson, Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The Tagged LOB Corpus: Users' manual. Bergen University*, Norway: ICAME, The Norwegian Computing Centre for the Humanities.

Adam Kilgarriff, Siva Reddy, Jan Pomikalek, and Avinesh PVS. 2010. A Corpus Factory for many languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Malta.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *Machine Translation Summit X*. pp.79-86.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. pp.230–237, Barcelona, Spain.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German Reference Corpus DeReKo: A primordial sample for linguistic research. In *Calzolari, N. et al. (eds.): Proceedings of the 7th conference on International Language Resources and Evaluation (LREC).* pp.1848–1854. Valletta, Malta: European Language Resources Association (ELRA).

Henry Kucera and Nelson Francis. 1967. *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conf. on Machine Learning*. pp.282–289. Morgan Kaufmann, San Francisco, CA.

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp.161-164. Jeju Island, Korea

Paul M. Lewis (ed.). 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. *Online version: http://www.ethnologue.com/.*

Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

Michal Marek, Pavel Pecina and Miroslav Spousta. 2007. Web page cleaning with conditional random fields. In *Proceedings of the Web as Corpus Workshop (WAC3), Cleaneval Session*, Louvain-la-Neuve, Belgium.

Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. 1993. Structural matching of parallel texts. In *31st Annual Meeting of the Association for Computational Linguistics: ACL-93*, pp.23–30.

Yuji Matsumoto, Kazuma Takaoka and Masayuki Asahara. 1999. ChaSen Morphological Analyzer version 2.4.0 User's Manual. NAIST Technical Report, Nara Institute of Science and Technology Technical Report 99009. *Retrieved on 07 Jan 2011 from http://sourceforge.jp/projects/chasen-legacy/docs/chasen-2.4.0-manual-en.pdf/en/1/chasen-2.4.0-manual-en.pdf.pdf*

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19 (2). pp.313--330 (Special Issue on Using Large Corpora).

Tony McEnery & Andrew Wilson. 2001. *Corpus Linguistics (2nd ed)*. Edinburgh : Edinburgh University Press.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to WordNet: An on-line lexical database (revised). *Technical Report Cognitive Science Laboratory (CSL) Report 43,* Princeton University, Princeton. Revised March 1993.

Roberto Navilgli. 2006. Consistent validation of manual and automatic sense annotations with the aid of semantic graphs. *Computatational Linguistic* 32(2), pp.273–281.

Paroo Nihilani. 1992. 'The International Computerized Corpus of English'. In *Anne Pakir (ed.) Words in a cultural context*. Singapore: UniPress. pp.84-88.

Cam Tu Nguyen, Trung Kien Nguyen, Xuan Hieu Phan, Le Minh Nguyen, and Quang Thuy Ha. 2006. Vietnamese Word Segmentation with CRFs and SVMs: An Investigation. In *Proceedings of The 20th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*. Wuhan, China.

Cam-Tu Nguyen, Xuan-Hieu Phan and Thu-Trang Nguyen. 2010. JVnTextPro: A Java-based Vietnamese Text Processing Tool. *Retrieved on 04 Feb 2011 from http://jvntextpro.sourceforge.net/*

Cam-Tu Nguyen and Xuan-Hieu Phan. 2007. JVnSegmenter: A Java-based Vietnamese Word Segmentation Tool. *Retrieved on 30 Jan 2011 from http://jvnsegmenter.sourceforge.net/*

Franz J. Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proc. of the 18th Int. Conf. on Computational Linguistics*, pp.1086-1090

Online OCR. 2009. Free Online OCR-convert scanned PDF and images to Word, JPEG to Word. *Available from http://www.onlineocr.net/*

Vincent Ooi .1997. Analysing the Singapore ICE corpus for lexicographic evidence. In *Magnus Ljung (ed.) Corpus-based studies in English*. Amsterdam: Rodopi. pp.245-260.

Pam Peter. 1987. Towards a corpus of Australian English. *International Computer Archive of Modern English Journal*, 11, pp.27-28

Jan Pomikalek, Pavel Rychly and Adam Kilgarriff. 2009. Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics. *Special Issue of Research in Computing Science* vol 41, Mexico City.

POSTECH, Intelligent Software Lab. 2008. Intelligent Software Lab. *Retrieved on 07 Mar 2011 from http://nlp.postech.ac.kr/DownLoad/k_api.html*

Desmond D. Putra, Abul Arfan and Ruli Manurung. 2008. Building an Indonesian WordNet. In *Proceedings of the 2nd International MALINDO Workshop.*

Randolph Quirk, Sidney Greenbaum, Geoffrey. Leech, Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer and Ian Roberts. 2008. Semantic annotation of clinical text: the CLEF corpus. In *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*. pp.19–26.

Kergrit Robkop, Sareewan Thoongsup, Thatsanee Charoenporn, Virach Sornlertlamvanich and Hitoshi Isahara. 2010. WNMS: Connecting the Distributed WordNet in the Case of Asian WordNet. In *The 5th International Conference of the Global WordNet Association (GWC-2010), Mumbai, India*

Xavier Roche. 2007. Httrack Website Copier - Offline Browser .*Retrieved Jan 30, 2011 from http://www.httrack.com/*

John M. Sinclair. 2005. Corpus and Text - Basic Principles. In *Developing Linguistic Corpora: a Guide to Good Practice, (ed.) M. Wynne*. Oxford: Oxbow Books: 1–16

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proceedings of RANLP'09*. pp.237–248. Borovets, Bulgaria

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. *Retrieved on 03 Jan 2011 from http://www.natcorp.ox.ac.uk/*

Kristina Toutanova,; Christopher D. Manning. 2000. "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger". In *Proceedings of J. SIGDAT*

Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-2000). pp.63–70.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A Conditional Random Field Word Segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing.*

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the Fourth SIGHAN Workshop*, Jeju Island, Korea.

T-reinhardt.ch. n.d. Free Online OCR. *Available from http://www.free-ocr.com*

Daniel Varga, Laszlo Nemeth, Peter Halacsy, Andras Kornai, Viktor Tron, and Viktor Nagy. 2005. Parallel corpora for medium density languages. *In Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*. pp.590–596. Borovets. Bulgaria.

VLSP Project (2006-2010). VLSP Project – Vietnamese Language Processing. *Retrieved on 02 Jan 2010 from http://vlsp.vietlp.org:8080/demo/?page=about*

Martin Volk and Yvonne Samuelsson, 2004, Bootstrapping parallel treebanks. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC) at COLING 2004*, Geneva, Switzerland

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proceedings of the second International Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

George K. Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin.

Appendix

Copyrights

Penn TreeBank II Tagset used by HunPos (Mitchell et al, 1993)

Tags	Grammatical Categories	POS Count
CC	Coordinating conjunction	3,591
CD	Cardinal number	1,345
DT	Determiner	7,528
EX	Existential there	67
FW	Foreign word	596
IN	Preposition or subordinating conjunction	9,597
JJ	Adjective	10,347
JJR	Adjective, comparative	186
JJS	Adjective, superlative	292
LS	List item marker	0
MD	Modal	649
NN	Noun, singular or mass	17,578
NNS	Noun, plural	6,362
NNP	Proper noun, singular	69
NNPS	Proper noun, plural	0
PDT	Predeterminer	39
POS	Possessive ending	1
PRP	Personal pronoun	1,163
PRP$	Possessive pronoun	962
RB	Adverb	2,693
RBR	Adverb, comparative	86
RBS	Adverb, superlative	113
RP	Particle	328
SYM	Symbol	8
TO	to	1,644
UH	Interjection	6
VB	Verb, base form	2,842
VBD	Verb, past tense	650
VBG	Verb, gerund or present participle	1,385
VBN	Verb, past participle	2,016
VBP	Verb, non-3rd person singular present	1,158
VBZ	Verb, 3rd person singular present	2,215
WDT	Wh-determiner	455
WP	Wh-pronoun	137
WP$	Possessive wh-pronoun	6
WRB	Wh-adverb	233
Total tokens inclusive of punctuations and symbols		**76,347**

Penn Chinese TreeBank Tagset used by Stanford Chinese POS tagger (Xia et al, 2000)

Tags	Grammatical Categories	POS Count
AD	Adverb	4,368
AS	Aspect marker	409
BA	把 in 把 ba construction	65
CC	Coordinating conjunction	1,692
CD	Cardinal number	1,348
CS	Subordinating conjunction	193
DEC	的 in a relative clause	2,025
DEG	Associative 的	2,024
DER	得 de in V-de construction V-de-R	7
DEV	地 before VP	62
DT	Determiner	1,164
ETC	For tokens with 等 or 等等	158
FW	Foreign words	1
IJ	Interjection	0
JJ	Other noun	1,681
LB	被 bei, in long bei construction	13
LC	Localizer	804
M	Measure word	1,448
MSP	other particle	168
NN	Common noun	17,801
NR	Proper noun	3,523
NT	Temporal noun	437
OD	Ordinal number	79
ON	Onomatopoeia ,	0
PP	Reposition excl. with 被 and 把	0
PN	Pronoun	1,607
PU	Punctuation	8,656
SB	被 bei, in short bei construction	48
SP	Sentence	68
VA	Predicative adjective	1,295
VC	是	709
VE	有 as the main verb	332
VV	Other verb	8,518
Total tokens inclusive of punctuations and symbols		60,703

JVnTextPro Tagset (Nguyen et al. 2010)

Tags	Grammatical Categories	Equivalence in Vietnamese Grammar	POS Count
A	Adjective	Tính từ	5,144
B	Loanwords	Từ mượn tiếng nước ngoài ví dụ Internet	0
C	Conjunction	Liên từ	4,297
E	Preposition	Giới từ	4,905
I	Interjection	Thán từ	15
L	Attributive	Dịnh từ	2,287
M	Numeral	Số từ	1,840
Mrk	Punctuations	Các dấu câu	6,111
P	Pronoun	Dại từ	873
R	Adjunct	Phụ từ	2,669
T	Particle, modal particle	Trợ từ, tiểu từ	459
V	Verb	Dộng từ	9,388
X	Un-known	Các từ không phân loại được	611
Y	Abbreviation	Từ viết tắt	2
N	Noun	Danh từ	14,194
Nb	Borrowed Noun	Danh từ có nguồn gốc nước ngoài	246
Nc	Classification Noun	Danh từ chỉ loại	1,150
Np	Personal Noun	Danh từ riêng	7,672
Ny	Abbreviated Noun	Các danh từ viết tắt	667
Nu	Unit Noun	Danh từ đơn vị	246
Total tokens inclusive of punctuations and symbols			62,646

MeCab-ipadic 2.7.0 tagset (Asahara and Matsumoto, 2003)

The MeCab-ipadic 2.7.0 POS categories are organized into hierarchies with the most basic categories as the first set of characters followed by a dash for the specific genre of sub-categories. And sometimes more than one set of POS is necessary to describe the current POS tag for the token; there will be a whitespace in between POS tags when more than one POS tag is necessary.

For example the POS tag, the token シンガポール *shingaporu* "Singapore" is annotated with the POS tag 名詞-固有名詞-地域-国 *meishi-koyu meishi-chiiki-koki* "noun-proper noun-place-country". 名詞-固有 *meishi-koyu* "noun-proper" refers to シンガポール *shingaporu* "Singapore" as a proper noun, and 名詞-地域-国 *meishi-chiiki-koki* "noun-place-country" refers to シンガポール *shingaporu* "Singapore" as a noun of a place and more specifically a country.

Due to the productivity of ipadic in annotating text, only the generic level of the POS will be presented in this dissertation (refer to Asahara and Matsumoto, 2003 for a detailed manual of ipadic 2.7.0). A column of sub-categorical POS tags will be included to draw parallel with the ones used by the English Penn TreeBank tagset.

Generic Categories	**No. of subcategories in ipadic 2.7.0**	**Example of Subcategorical POS**	**Closest fit to English Penn Tree Bank tagset**	**POS Count**
名詞 (Nouns)	31	名詞-一般 (noun-common)	NN	
		名詞-固有 (noun-proper)	NNP	
		名詞-数 (noun-numeral)	CD /OD	
		名詞-サ変接続 (noun-verbal)	-	32,685
接頭詞 (Prefix)	4	接頭詞-名詞接続 (prefix-nominal)	-	
		接頭詞-数接続 (prefix-numerical)	-	
		接頭詞-動詞接続 (prefix-verbal)	-	
		接頭詞-形容詞接続 (prefix-adjectival)	-	754
動詞 (Verb)	34	基本形 (basic form)	VB	
		動詞-非自立 (aux form)	MD	
		動詞-接尾 (verbal suffix)	-	9,531
形容詞 (Adjective)	10	形容詞-自立 (main adjective)	JJ	
		形容詞-非自立 (bounded/sub adjective)	JJ	
		形容詞-接尾 (adjectival suffix)	-	931

副詞 (Adverb)	2	副詞-一般 (misc adverb) (i.e. adverbs that can be segmented into one unit and where adnominal modification is not possible)	RB	737
		副詞-助詞類接続 (particle adverb) (i.e. Adverbs that can be followed by particle e.g. 「の」 「は」 「に」 「な」 「する」 「だ」 etc.)	RB	
助詞 (Particle)	13	助詞-格助詞-一般 (case particle)	-	20,786
		助詞-接続助詞 (conjunctive particle)	-	
		助詞-副助詞 (adverbial particle)	-	
		助詞-終助詞 (final particle)	-	
助動詞 (Aux verb)	14	助動詞 不変化型 (non-inflectional aux)	MD	6,200
		助動詞 形容詞・イ段 (inflectional adjectival aux)	MD	
記号 (Symbol)	7	記号-句点 (symbol-comma)	,	11,280
		記号-読点 (symbol-period)	.	
感動詞 (Exclamation)	1	感動詞 (exclamation)	-	17
接続詞 (Conjunction)	1	接続詞 (conjunction)	CC	375
フィラー (Filler)	1	フィラー (filler) i.e. Aizuchi that occurs during a conversation or sounds inserted as filler e.g. 「あの」 a-no 「えと」 e-to	-	23
連体詞 (Adnominal)	1	連体詞 (adnominal) i.e. Words that only have noun-modifying forms.	-	753
その他-間投 (Other-interjection	1	その他-間投 (other-interjection)) ie. words that are hard to classify as noun-suffixes or sentence-final particles. e.g. 「(だ)ァ」 (da)a	-	5
Total tokens inclusive of punctuations and symbols				84,077

POSTAG/Sejong tagset (Lee et al, 2002)

Tags	Grammatical Categories	POS Count
EC	Conjunctive endings	4,234
EF	Terminal endings	2,588
EP	Non-terminal endings	767
ETM	Adnominalisers	5,660
ETN	Nominalisers	379
IC	Interjections	9
JC	Conjunctive case particles	837
JKB	Adverbial case particles	3,297
JKV	Vocative case particles	4
JKC	Complementative case particles	22
JKG	Genitive case particles	1,301
JKO	Accusative case particles	2,733
JKQ	Quotative case particles	20
JKS	Nominative case particles	1,244
JX	Auxiliary particles	2,514
MAG	General adverbs	1,543
MAJ	Conjunctive adverbs	229
MM	Adnominals	983
NNB	Bound nouns	2010
NNG	General nouns	19,752
NNP	Proper nouns	2,786
NP	Pronouns	482

Tags	Grammatical Categories	POS Count
NR	Numerals	147
SE	Ellipsis symbols	0
SF	., !, ?	1
SH	Words in Chinese characters	4
SL	Foreign words	3,150
SN	Numbers	606
SO	~	19
SP	,, :, /, ·	2,453
SS	Quotation marks, brackets, dash	2,360
SW	Logical and mathematical, currency symbols	174
VA	Adjectives	2,264
VCN	Negative copula	59
VCP	Positive copula	1,544
VV	Verbs	4,558
VX	Auxiliary predicates	1,354
XPN	Nominal prefixes	90
XR	Root	1,016
XS	Suffixes	0
XSA	Adjectival derivational suffixes	1,129
XSB	Adverbial derivational suffixes	0
XSN	Nominal derivational suffixes	1,210
XSV	Verbal derivational suffixes	1,823
Total tokens inclusive of punctuations and symbols		77,355

Sejong-Shell Script

```
#!/bin/bash -x
################################################################################
## Sejong-Shell is a script to call POSTAG/Sejong tagger on Unix machine
## because POSTAG/Sejong is only usable in Korean Microsoft Windows environment
## the original POSTAG/Sejong can be downloaded from
## http://isoft.postech.ac.kr/Course/CS730b/2005/index.html
##
## Sejong-Shell is dependent on  WINdows Emulator.
## The WINE program can be downloaded
## from http://www.winehq.org/download/
##
## The shell scripts accepts the input text file from one directory and
## outputs the tagged files into another while retaining the filename
################################################################################

cd <source_file_dir>
#<source_file_dir> is the directory that saves the textfiles that needs tagging
for file in `dir -d *`
do
        echo $file
        cp <source-file_dir>/"$file" <POSTAG-Sejong_dir>/input.txt
        # <POSTAG-Sejong_dir> refers to the directory where
        # the pos-tagger is saved
        wine start /Unix "<POSTAG-Sejong_dir>/sjTaggerInteg.exe"
        sleep 30
        # this is necessary so that the file from the current loop won't be
        # overlapping with the next, do increase the time for sleep if the file
        # is large and needs more than 30sec for POSTAG/Sejong to tag
        cp <POSTAG-Sejong_dir>/output.txt <target-file_dir>/"$file"
        # <target-file_dir> is where you want the output files to be stored
done
```

Glossary of Computational Linguistic, NLP and Programming Terminologies

Conditional Random Fields (CRF) is a statistical model that is used to parse sentences as sequential tokens. In the case of POS tagging and word segmentation, other statistical model treats tokens as a single entity and calculate the probability of the correct tag/segment of the particular token based on previous tokens. For CRF, it treats the whole chunk of tokens as an entity and calculate the most probably chunk that should appear in the given sequence of tokens (for more details refer to Lafferty et al, 2001)

EUC-KR (Extended Unix Code - KoRean) is a multibyte character encoding for Japanese, Korean and simplified Chinese. The POSTAG/Sejong tagger used in this project only accepts EUC-KR inputs and outputs also in EUC-KR

F-score is the mathematically harmonic score from a program's precision and recall. It is a mathematical evaluation to calculate the effectiveness of the computer programs, the higher score equates to better program (refer to van Rijsbergen (1979) for details of mathematical evaluation of computer programs)

HMM (Hidden Markov Model) is a pattern recognition model that calculates the probability of the hidden (i.e. "next") finite state based on the probability of the transition from the previous state to the hidden state. In POS tagging tasks, the states refer to token and the probability calculations are based on how probable is a particular POS tag appearance given previous series/chains of POS tag. A token will be tagged with the POS from the highest probability of a given chain of POS (Ephraim and Merhav, 2002).

ISO-8859-1 (International Organization for Standardization-8859-1) is a single-byte character encoding, commonly used for European languages with Latin alphabet. The HunPos tagger used in this project outputs in this encoding.

MaxEnt (Maximum Entropy) refers to the Bayesian probability distribution which best represents the current state of knowledge. In POS tagging task for a sentence, the MaxEnt model calculates the probabilities of the every possible POS for each tokens given the previous POS (i.e. the Bayesian probability), the POS tag that scores the highest probability will be selected as the resultant POS tags for the sentence (Guiasu and Shenitzer, 1985).

MEMM (MaxEnt Markov Model) combines the MaxEnt and HMM by calculating the MaxEnt probability of a Markov chain. In MaxEnt model, the POS tag are considered as individual tokens that is only dependent of the previous POS tag (Bayesian probability), in MEMM the individual tokens are dependent on previous chains of POS tag, and probabilities are assigned to the chains and not the individual tags (Toutanova and Manning, 2000).

HTML::TreeBuilder – is a parser module for perl code that builds a HyperText Markup Language (HTML) syntax. The module allows the computer to read content bounded by HyperText markups; it easy access to content bounded by <> and </> tags. It is used by this project for cleaning and extracting text data from the raw HTML files. (Module documentation can be referred to from http://search.cpan.org/~jfearn/HTML-Tree-4.1/lib/HTML/TreeBuilder.pm)

Httrack – The corpus project used the httrack command `httrack http://www,yoursingapore.com -o +*.yoursingapore.com/content/traveller/*/* -p1` to download the html files from the *www.yoursingapore.com* website. Below explains the options used in downloading the files:

httrack [URL to download] [-option] [+ to include the html pages from the URL] [-p1]

`-o`	request custom options to be used in downloading
`+*URL`	to include these URL in download
`/*/*`	the regex used to download webpage in different languages (e.g. /en/zouk.html for English; /ja/zouk.html for Japanese)
`-p1`	this is the option to only download html files

Sed – is the **s**tream **ed**itor command that read and lines from `stdin` (standard input stream from the terminal) according to a pattern specified by the user using basic regular expression (sed documentation: http://www.gnu.org/software/sed/manual/sed.html). The `sed` command `$sed '/^$/d' -i *.txt` was used to remove empty lines from the extracted text files in the cleaning task. A brief explanation of the regex used as described in the cleaning section:

`/`	refers to a start of a regex
`^`	refers to the beginning line
`$`	refers to the end of line
`/d`	is the option to delete the line if it matches the regex
`*.txt`	requires the command to be iterated for any textfiles in the directory/folder

SVM (Support Vector Machine) – Given a set of input data, the standard SVM predicts two possible classes that each input can take. In POS tagging tasks, each token is given two possible POS tags and the statistical model recursively tagged the same texts and each round of tagging the probabilities of the two tags on each token widen and eventually it reaches a threshold where only one tag will be deemed as appropriate for the token (Cortes and Vapnik, 1995).

UTF-8 (Universal character set Transformation Format - 8 bit) is the designate encoding used for this project. All languages available in NTU-MC is UTF-8 compatible.

Viterbi is a dynamic programming algorithm to predict the most probable hidden states in a sequence of observed events. In the case of POS tagging, it refers to guess the next POS tag through calculating the most probable (mathematically) POS tag will be correct given the previous POS tag, by recursively calculating probabilities of the best possible tag for each token, a Viterbi path is calculated. Eventually the current tag that the program uses to tag a token is based on the calculated Viterbi path (Forney, 1973).